

Completeness II: A signal-to-noise approach for completeness estimators applied to galaxy magnitude-redshift surveys

Luís Teodoro^{1,3*}, Russell Johnston^{2,3†} and Martin Hendry^{3‡}

¹*ELORET Corp., Space Science and Astrobiology Division, MS: 245-3, NASA Ames Research Center, Moffett Field, CA 94035-1000, USA*

²*Department of Physics, University of Western Cape, Belville, Cape Town, South Africa*

³*Department of Physics and Astronomy, Kelvin Building, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK*

24 May 2010

ABSTRACT

This is the second paper in our completeness series which addresses some of the issues raised in the previous article by Johnston et al. (2007) in which we developed statistical tests for assessing the completeness in apparent magnitude of magnitude-redshift surveys defined by two flux limits. The statistics, T_c and T_v , associated with these tests are non-parametric and defined in terms of the observed cumulative distribution function of sources; they represent powerful tools for identifying the *true* flux limit and/or characterising systematic errors in magnitude-redshift data.

In this paper we present a new approach to constructing these estimators that resembles an “adaptive smoothing” procedure – i.e. by seeking to maintain the same amount the information, as measured by the signal-to-noise ratio, allocated to each galaxy. For consistency with our previous work, we apply our improved estimators to the Millennium Galaxy Catalogue (MGC) and the Two Degree Field Galaxy Redshift Survey (2dFGRS) data, and demonstrate that one needs to use a s/n appropriately tailored for each individual catalogue to optimise the performance of the completeness estimators. Furthermore, unless such an adaptive procedure is employed, the assessment of completeness may result in a spurious outcome if one uses other estimators present in the literature which have not been designed taking into account “shot noise” due to sampling.

Key words: Cosmology: methods: data analysis – methods: statistical – astronomical bases: miscellaneous – galaxies: redshift surveys – galaxies: large-scale structure of Universe.

1 INTRODUCTION

In recent years the statistical analysis of galaxy redshift surveys has played a central role in cosmology, yielding stringent constraints on the parameters of both the underlying cosmological world model and on the clustering properties of galaxies as a function of redshift, environment and morphological type. However, both tasks are hampered by observational selection effects – due to e.g. detection limits in apparent magnitude, colour, surface brightness or some combination thereof. A wide range of statistical tools has been developed to identify, characterise – and hopefully to remove – the impact of observational selection effects from magnitude-redshift surveys. Presently, we have the initial data release from the WiggleZ Dark Energy Survey (Drinkwater 2010), which will attempt to measure the baryon acoustic oscillation (BAO) scale to within 2% from 240,000 emission line galaxies. There also has also been the

zCOSMOS survey (Lilly 2009; Zucca et al. 2009) that is exploring galaxy evolution through the role of environment at high redshift in the range $1.5 \lesssim z \lesssim 3.0$. To achieve such high precision in these measurements will require accurate understanding of the selection and, particularly with zCOSMOS, luminosity functions.

To fully understand the statistical properties of the aforementioned selection function it is crucial that we understand the role of completeness in apparent magnitude – meaning that all galaxies brighter than some specified limiting apparent magnitude (or, as is pertinent to this paper, with apparent magnitudes lying between some specified bright and faint limiting values) have been observed. A classical test for completeness in apparent magnitude is to analyse the variation in galaxy number counts as a function of the adopted limiting apparent magnitude (Hubble 1926). This test, which presupposes that the galaxy population does not evolve with time and is homogeneously distributed in space, is however not very efficient. More specifically, it is difficult to decide in practice whether deviations from the expected galaxy number count are indeed an effect of incompleteness in apparent magnitude, or are instead due to galaxy clustering and/or evolution of the galaxy lu-

* luis@astro.gla.ac.uk

† rjohnston@uwc.ac.za

‡ martin@astro.gla.ac.uk

minosity function – or indeed created by incomplete sampling in apparent magnitude. Of course in designing a completeness test one can also make use of distance information via galaxy redshifts; the still widely used and well-known V/V_{\max} test of Schmidt (1968) does this, and considers – for a specified magnitude limit – the ratio of two volumes: the volume of a sphere of radius equal to the actual distance of observed galaxy, divided by the volume of a sphere of radius equal to the *maximum* distance at which the galaxy would be observable – i.e. at the apparent magnitude limit. It follows that – for a non-evolving, homogeneous distribution of galaxies – the expected value and variance of V/V_{\max} are equal to $1/2$ and $1/12$ respectively. The V/V_{\max} test has been used to assess the completeness of magnitude-redshift samples (see for example Hudson and Lynden-Bell 1991), but unfortunately it suffers from the same major drawbacks as the Hubble test based on galaxy number counts: it is difficult to interpret whether any significant measured departure from the expected value of V/V_{\max} is due to incompleteness or to clustering and evolutionary effects.

In a seminal paper, Efron and Petrosian (1992) (hereafter EP92) introduced a powerful new approach to analysing magnitude-redshift surveys that drew on concepts developed in the so-called C -method of Lynden-Bell (1971) for constructing galaxy LFs. EP92 proposed a non-parametric permutation test for the independence of the spatial and luminosity distributions of galaxies in a magnitude-limited sample, which required no assumptions concerning the parametric form of both the spatial distribution and the galaxy luminosity function. They applied this test to a quasar sample, with an assumed apparent magnitude limit, in order to robustly estimate the parameters characterising the luminosity distance-redshift relation of the quasars (see also Efron and Petrosian 1999).

Rauzy (2001) (hereafter R01) noted that the essential ideas of EP92 could be straightforwardly adapted and extended to turn their non-parametric test of the cosmological model into a non-parametric test of the assumption of a magnitude-limited sample – thus developing a simple but powerful tool for assessing the magnitude completeness of magnitude-redshift surveys. As was the case with EP – and unlike the Hubble number counts or V/V_{\max} tests – the Rauzy test statistic, T_c , requires no assumption about the spatial homogeneity of the galaxy distribution. Moreover, it also requires no knowledge of the parametric form of the galaxy luminosity function. On the other hand, the Rauzy test was formulated only for the case of a sharp, faint apparent magnitude limit.

Johnston et al. (2007) (hereafter JTH07) discussed the advantages of the T_c statistic over standard completeness tests and extended its use to data that is characterised by both a faint and bright magnitude limit. Moreover, they introduced a new variant statistic, called T_v , constructed using the sampled cumulative distance modulus, Z , distribution that retains similar properties to those of T_c i.e. being independent of the spatial distribution of galaxies. By sampling the data in this way, the T_v statistic amounted to a much improved differential version of the widely used V/V_{\max} test (which assumes spatial homogeneity). JTH07 applied their completeness test to three major redshift surveys: the Millennium Galaxy Catalogue (MGC) (e.g. Liske et al. 2003; Cross et al. 2004), the Two Degree Field Galaxy Redshift Survey (2dFGRS) (e.g. Colless 2001), and a Sloan Digital Sky Survey - Early Types (SDSS-ET) (e.g. Bernardi 2003) sample. They concluded that all three surveys were complete in apparent magnitude up to their respective published magnitude limits. In the case of the 2dFGRS survey data, however, they showed that one is first required to adopt a secondary bright apparent magnitude limit – i.e. applying the JTH07 generalisation.

Application of the JTH07 generalised completeness test to these three surveys led us to consider two crucial effects that, if not accounted for correctly, could lead to wrong statistical conclusions concerning determination of the *true* completeness limits. In rough terms, the basic construction of the T_c and T_v statistics proceeds by identifying volume-limited subsamples associated with each individual galaxy in the catalogue. In the design of the original Rauzy completeness test, where one is only concerned with the faint apparent magnitude limit, these volume-limited subsamples were uniquely defined and thus could be allowed to grow such that a maximised sampling of the data was achieved. With the introduction of a secondary bright limit (as shown in Figure 1) the size of each volume-limited subsample is no longer unique. This leads to the obvious question: how should one optimally define each subsample?

In studying the distribution of galaxies in the (M, Z) plane we are seeking to understand the underlying luminosity function of a given population of galaxies, as well as the manner in which that function is sampled. *To do so we are, of course, inevitably limited to inferences drawn from a finite number of galaxies.* This makes the inference process in principle susceptible to shot-noise and thus, if our estimators are constructed from subsamples which are too sparsely populated, might lead to spurious results concerning the global properties of the data-set. In this paper we therefore propose to optimise and extend our current methodology by invoking a well-established and objective criterion: we construct our completeness estimators so as to maximize their local signal-to-noise ratio.

The format of this paper will be as follows. In § 2 we revisit the main points underpinning the construction of the JTH07 T_c and T_v statistics. In § 3 we then explore the adverse consequences that can arise if the JTH07 method is applied without properly accounting for the impact of sparse sampling. For this exploration we use the Millennium Galaxy Catalogue (MGC) and the Two Degree Field Galaxy Redshift Surveys (2dFGRS), as already studied in JTH07, for purely illustrative purposes. This then leads us, in § 4, to propose an optimisation technique that is the first step towards circumventing these issues. In § 5 we introduce as a sampling threshold a direct measurement of the signal-to-noise (s/n) of our sampling technique, and demonstrate how this can be implemented. In § 6 we then discuss our conclusions and future work.

2 THE ‘SEPARABILITY’ ASSUMPTION AND STATISTICAL FRAMEWORK

We recall that the fundamental assumption of our method – also referred to as ‘separability’ – is that the luminosity function of the galaxy distribution is not dependent on the three-dimensional redshift space positions $\mathbf{z} = (z, l, b)$ of the galaxies, where (l, b) are galactic directional coordinates. Although this is a rather restrictive assumption it underlies most of the traditional completeness tests in the literature. The corrected distance modulus Z is defined as,

$$Z \equiv m - M = \mu(z) + k_{\text{corr}} + e_{\text{corr}} + A_g(l, b), \quad (1)$$

where k_{corr} and e_{corr} are the k -correction and evolutionary correction, respectively, $\mu(z)$ is the distance modulus at redshift z and $A_g(l, b)$ is an extinction correction dependent on galactic coordinates. For simplicity we are marginalising over the galactic directional coordinates.

In assuming separability the joint probability density in absolute magnitude and corrected distance modulus can therefore be

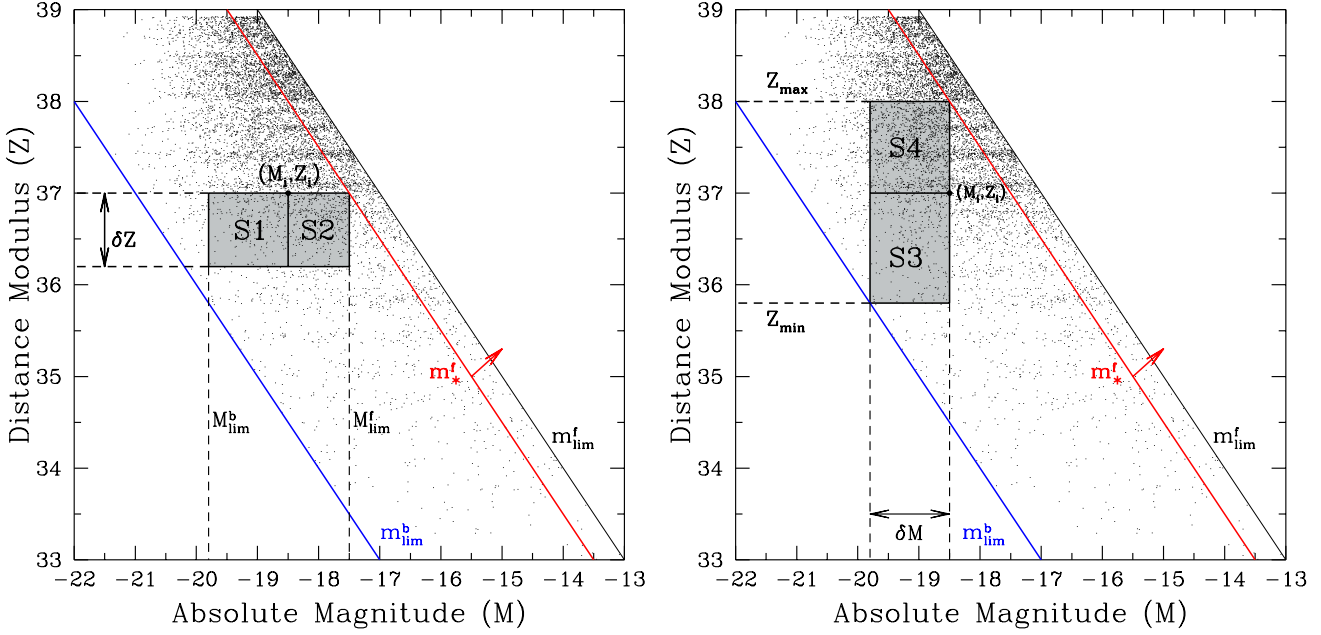


Figure 1. Diagram illustrating the construction of the rectangular regions S_1, S_2 (left), and S_3, S_4 (right) which are defined for the random variables, ζ_i and τ_i respectively, for a typical galaxy at (M_i, Z_i) . The left hand panel shows the construction of the regions S_1 and S_2 with the inclusion of bright and faint limits m_{lim}^b and m_{lim}^f , respectively. These regions are uniquely defined for a ‘slice’ of specified width, δZ , in distance modulus, and a ‘trial’ faint limit m_{*}^f . The right hand panel illustrates the construction of the rectangular regions S_3 and S_4 for τ_i . These regions are uniquely defined for a ‘slice’ of specified width, δM , in absolute magnitude and a ‘trial’ faint limit m_{*}^f .

written as,

$$dP = [h(Z) dZ] [f(M) dM] \theta(m_{\text{lim}}^f - m) \theta(m - m_{\text{lim}}^b), \quad (2)$$

where $f(M)$ and $h(Z)$ are the probability density function of M and Z , respectively, and θ is the Heaviside or ‘step’ function defined as,

$$\theta(x) \equiv \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \quad (3)$$

Thus for each object i present in a catalogue we define the random variables ζ_i and τ_i for the statistics T_c and T_v respectively¹ (for a detailed discussion see JTH07),

$$\begin{aligned} \zeta_i &= \frac{F(M_i) - F[M_{\text{lim}}^b(Z_i - \delta Z)]}{F[M_{\text{lim}}^f(Z_i)] - F[M_{\text{lim}}^b(Z_i - \delta Z)]} \\ &= \frac{n(S_1)}{n(S_1 \cup S_2)} = \frac{r_i}{n_i + 1}, \end{aligned} \quad (4)$$

and

$$\begin{aligned} \tau_i &= \frac{H(Z_i) - H[Z_{\text{min}}^b(M_i - \delta M)]}{H[Z_{\text{max}}^f(M_i)] - H[Z_{\text{min}}^b(M_i - \delta M)]} \\ &= \frac{n(S_3)}{n(S_3 \cup S_4)} = \frac{q_i}{t_i + 1}, \end{aligned} \quad (5)$$

where r_i denotes the number of galaxies belonging to region S_1 ,

¹ Briefly, T_c and T_v are defined as

$$T_c = \sum_{i=1}^{N_{\text{gal}}} \frac{\zeta_i - 1/2}{[\text{Var}(\zeta_i)]^{1/2}}, \quad \text{and} \quad T_v = \sum_{i=1}^{N_{\text{gal}}} \frac{\tau_i - 1/2}{[\text{Var}(\tau_i)]^{1/2}},$$

respectively.

n_i the number of galaxies belonging to $S_1 \cup S_2$, q_i the number of galaxies belonging to S_3 , and t_i the number of galaxies belonging to $S_3 \cup S_4$. Figure 1 illustrates the construction of the rectangular regions S_1, S_2, S_3 and S_4 as well as the meaning and definition of the slices in magnitude, δZ , and distance modulus, δM . It should be mentioned that r_i was also the notation used in EP92 to denote the *rank* of the object i when galaxies are sorted by magnitude.

Essentially, the key to the JTH07 extension lay in the introduction of these fixed ‘slice’ widths δZ for ζ_i and δM for τ_i . Fixing these widths to a predetermined value allows the construction of unique, separable regions in Equations 4 and 5 within any doubly truncated survey i.e. for a survey with well defined bright and faint apparent magnitude limits.

However, the choice of the δZ and δM widths is essentially arbitrary, and one might wish to consider applying different ‘trial’ widths depending on the properties of the data set under study. JTH07 briefly discussed this point, and noted that by varying the widths in this manner two distinct effects for the determination of the true m_{lim}^f were revealed:

- For very small values of δZ and δM the respective T_c and T_v statistics will be dominated by what we may term ‘shot-noise’ (since the rectangular regions they identify are extremely sparsely sampled); this makes the process of drawing significant conclusions regarding nature of the true faint apparent magnitude limit impossible.

- Conversely, when the values of δZ and δM are taken to be very large, then for data-sets that are not well described by a sharp m_{lim} one appears to observe a range of possible values for the *true* faint magnitude limit.

We will illustrate in more detail the manifestation of these two effects in the following section.

3 CONSEQUENCES OF SPARSE SAMPLING

For continuity (and illustrative purposes) we revisit the Millennium Galaxy Catalogue (MGC), the Two Degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey Early Types (SDSS-ET) samples as used previously in JTH07. Please refer to this paper for survey description and sample selection.

3.1 ‘Shot-noise’ dominated sampling

In this section we examine more closely the consequences of sparse sampling issues in the construction of the random variables, ζ_i and τ_i , for the statistics T_c and T_v respectively.

In Figure 2 we have applied the JTH07 T_c and T_v estimators to the SDSS-ET (upper panel), MGC (middle panel) and 2dFGRS (lower panel) for selected values of δZ and δM . (Both δZ and δM are defined in Figure 1). For the SDSS-ET we observe that the T_c and T_v curves corresponding to respective widths of δZ and $\delta M = 0.001$ and 0.01 fluctuate within the $|3\sigma|$ limits for each m_* between the survey limits of $14.5 < m_{\text{lim}} < 17.45$ (as one would expect for a complete sample, following EP92, R01 and JTH07). However, contrary to the expectations of those earlier papers, as m_* moves beyond the published faint limit of the survey, the T_c and T_v curves drop slightly and then flatten (or ‘flat-line’) inside $-3\sigma < T_c, T_v < 3\sigma$ regions, instead of dropping sharply below the -3σ level. Similar results are seen with MGC at δZ and $\delta M = 0.01$ and 2dF up to δZ and $\delta M \approx 0.02$. As we move to increasingly larger values of δZ and δM , as shown in Figure 2, the T_c and T_v curves continue to ‘flat-line’ beyond the magnitude limit, but now do so at a value of the statistic which lies increasingly below -3σ .

This so-called ‘flat-lining’ effect can essentially be used as a means of identifying the ‘shot-noise’ level for a given width of δZ and δM – i.e. the value of δZ and δM less than which the sampling becomes too sparse to allow the magnitude limit to be reliably estimated. Understanding precisely why this ‘flat-lining’ happens becomes quite straightforward when one considers carefully what are the contributing factors: the number of objects in the catalogue and the range in apparent magnitude of the survey. The effect is illustrated in detail in Figure 3. The left panel shows the now familiar M - Z distribution with the red diagonal lines representing the faint apparent magnitude limit m_{lim}^f and our adopted bright limit m_{lim}^b . The main feature of this plot is the narrow red, blue and green rectangles which actually delineate the T_c regions S_1 and S_2 for a galaxy at (M_i, Z_i) with $\delta Z = 0.001, 0.008$ and 0.02 respectively. (Here we are considering a trial m_* equal to the survey limit i.e. $m_{\text{lim}}^f = 19.45$ mag.). Since these rectangular ‘strips’ represent such a tiny fraction of the M - Z distribution they can barely be separated in the main diagram. The left panel, therefore, also shows a close-up of this particular region, where the distinctive coloured areas are now clearly defined. (Note that, because of the very narrow range of distance moduli considered in this close-up, the apparent magnitude limit appears essentially as a vertical line). The right panel of Figure 3 represents, for the same galaxy at (M_i, Z_i) , the equivalent T_v construction with $\delta M = 0.001, 0.008$ and 0.02 respectively – with again the different coloured regions also shown in extreme close-up.

What is immediately apparent for both the T_c and T_v statistics is the very small number of galaxies that populate the rectangular regions for these small values of δZ and δM . In particular, it is clear that as m_* is increased beyond the true value m_{lim}^f , no further galaxies will be added to the subsets S_2 (for T_c) and S_4 (for T_v). By considering Equations 4 and 5, it then follows that the T_c and

T_v statistics will remain constant for larger values of m_* – which explains the ‘flat-lining’ effect seen in Figure 2.

The pattern which was apparent in Figure 2, whereby the ‘flat-lining’ effect occurred at progressively lower values of T_c and T_v as the widths of δZ and δM were increased, can be extended to the limiting case that corresponds to the original Rauzy (R01) completeness test – where the absence of a bright apparent magnitude limit means that there is in principle no limit to the height of the constructed regions. However, since we are dealing with a flux-limited catalogue that contains a finite number of galaxies, we can expect that *ultimately* the ‘flat-lining’ effect will become apparent for the R01 completeness test too, if we consider a sufficiently faint trial value of m_* . This effect is indeed seen in Figure 4, albeit for a value of T_c and T_v that lies enormously below the characteristic $3 - \sigma$ level which one might choose to identify as the value of the statistic indicating the true apparent magnitude limit.

In summary, then, we can understand the ‘flat-lining’ effect as a direct consequence of the very sparse sampling which occurs for small values of δZ and δM . A suitable choice for the width of δZ and δM can then be taken to be the values for which the onset of the ‘flat-lining’ effect *only* occurs when the test statistics T_c and T_v have already dropped to $3 - \sigma$ below their expected value, when the trial apparent magnitude limit is equal to the true value m_{lim}^f .

3.2 Variation in m_{lim}

We now briefly consider the apparent variation in the value of m_{lim} determined, resulting from the adoption of larger values of δZ and δM for a survey that is doubly truncated by a bright and faint magnitude limit. If we consider once again Figure 2 we can see that for both SDSS-ET and MGC as we move to larger values of δZ and δM our ability to determine correctly the true completeness limit of the survey is unaffected. However, we can quite clearly see in the case of the 2dFGRS on the lower panel that, as δZ and δM increases to, and beyond, the point at which the test statistics systematically fall below their -3σ level (which we adopt to indicate the true apparent magnitude limit), the value of $m_{\text{lim}}^{\text{true}}$ also varies with the values of δZ and δM adopted. In the range of values we are considering in this example, $0.002 \lesssim \delta Z, \delta M < 0.5$ we actually observe a corresponding range of m_{lim} from $19.0 \lesssim m_{\text{lim}} \lesssim 19.4$.

This variation in the ‘true’ magnitude limit inferred for a survey is rather unsatisfactory, and would somewhat defeat the purpose of the original Rauzy completeness test: to provide a robust, non-parametric and objective method for independently validating the magnitude completeness of a given survey. It underlines the importance of optimising the performance of our test statistics – an issue which we consider in more detail in the following sections.

4 EXPRESSIONS FOR THE SIGNAL-TO-NOISE OF OUR ESTIMATORS

In this section we now consider how the estimators ζ_i and τ_i are constructed, and in particular how they will be affected by random sampling fluctuations, in order to gain insight on how they might be optimised. This will essentially involve computing a measure of the s/n on the sampled ζ_i and τ_i , and how those variables are affected by fluctuations in the number of galaxies sampled in the regions S_1 to S_4 . For the moment let us consider ζ_i only. If we assume that the survey galaxies are sampled according to a Poisson distribution then we can derive an expression for the Poisson (or shot) noise

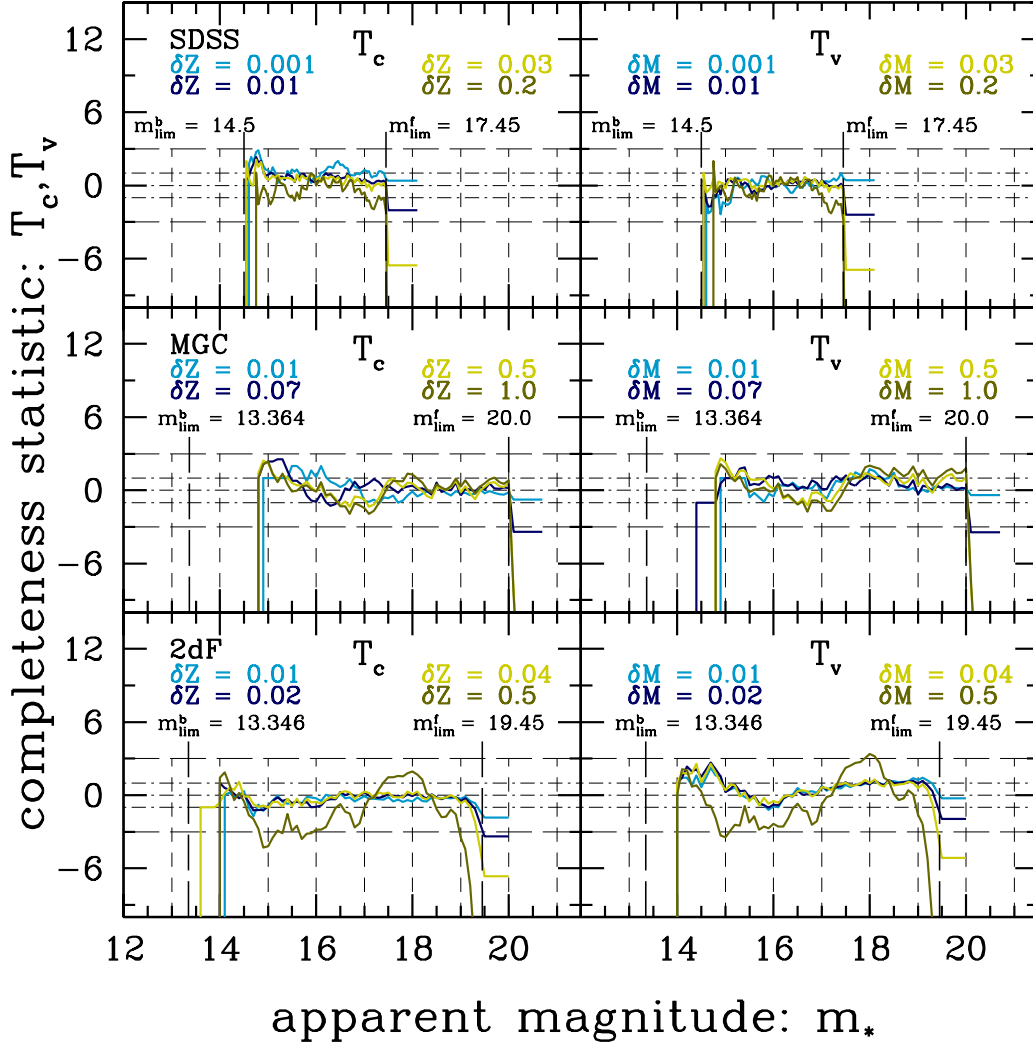


Figure 2. T_c and T_v Results for the SDSS-ET (upper panels), MGC (middle panels) and 2dF (lower panels) applying the JTH07 method for varying values of δZ and δM , where we can observe the transition between ‘shot-noise’ dominated sampling and signal dominated sampling. We can define this transition to occur at the point where the width of δZ or δM , for T_c and T_v respectively, is sufficiently large that the appropriate statistic drops to the -3σ level at the faint magnitude limit of the survey. This occurs at values of δZ and $\delta M \gtrsim 0.01$ for SDSS, $\gtrsim 0.07$ for MGC and $\gtrsim 0.02$ for 2dF.

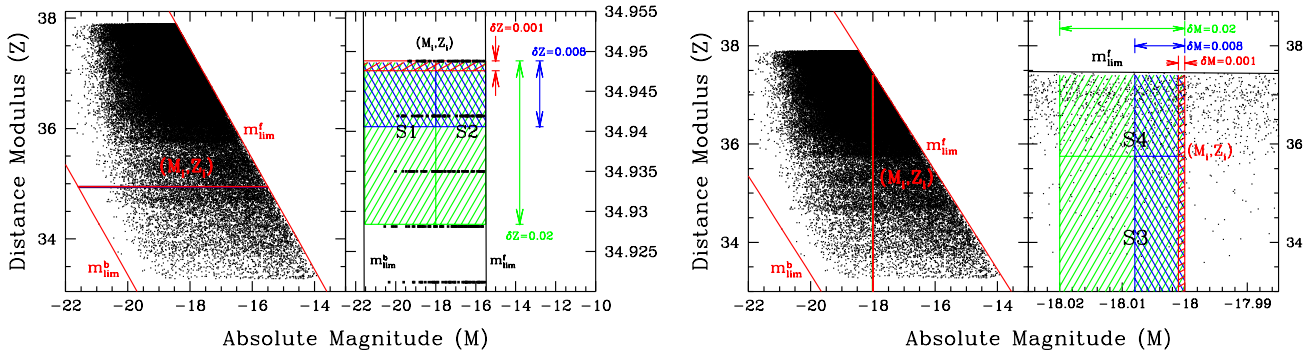


Figure 3. Schematic illustrating the cause of the ‘flat-lining’ effect (within the $[3\sigma]$ confidence limits) observed for small values of δZ and δM . The left hand panel shows the (M, Z) distribution for the 2dFGRS with the faint apparent magnitude limit m_{lim}^f and our adopted bright limit, m_{lim}^b indicated as red diagonal lines. The left hand plot considers the ζ_i construction for a galaxy at (M_i, Z_i) , with $\delta Z = 0.001, 0.008$ and 0.02 . The right hand plot zooms in to allow us to see the three distinct regions created as the size of δZ , and the relative number of galaxies contained therein, increases. Similarly, the right hand panel shows the corresponding τ_i construction for a galaxy at (M_i, Z_i) , for $\delta M = 0.001, 0.008$ and 0.02 .

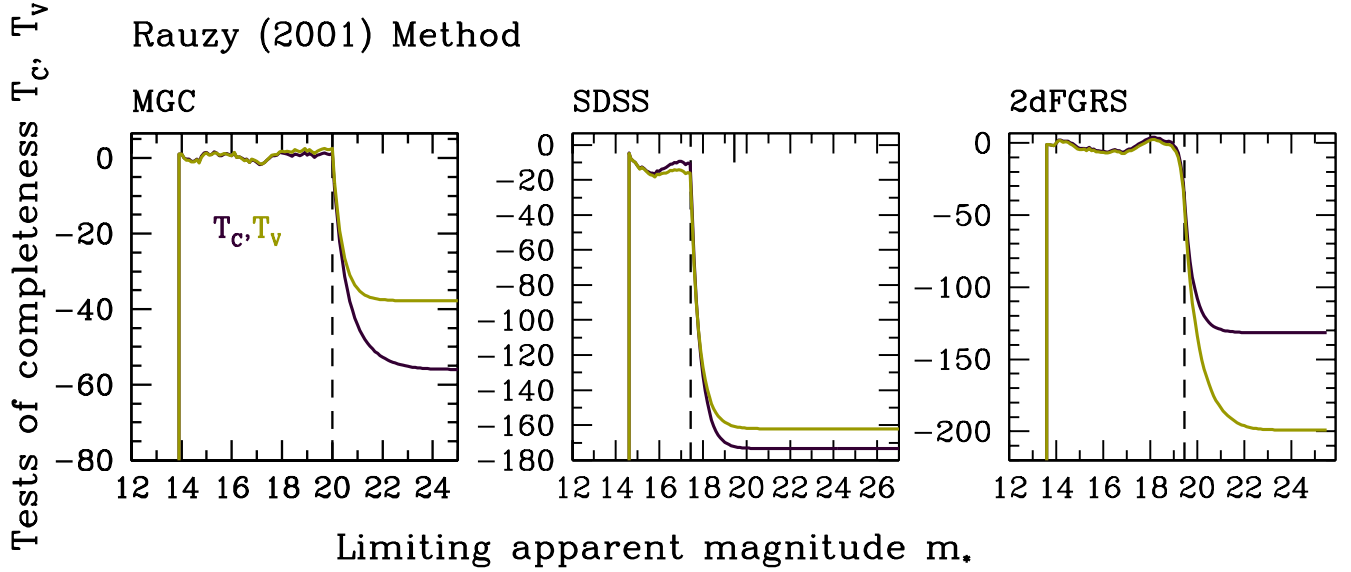


Figure 4. T_c and T_v plots for 2dFGRS (left) and MGC (right). Here we apply the R01 method, where the rectangular regions in Figure 1 are allowed to grow to their maximum size when accounting for m_{lim}^f only. We can see in both panels that if one allows m_* to pass far enough beyond the magnitude limit of the survey, the ‘flat-lining’ effect will eventually dominate, albeit for extremely negative values of T_c and T_v .

associated with ζ_i by applying simple perturbation theory. In this case Equation 4 then becomes

$$\delta\zeta_i = \frac{\delta r_i(n_i + 1) - r_i\delta(n_i + 1)}{(n_i + 1)^2}. \quad (6)$$

To take into account the cross-terms we square Equation 6 to get,

$$(\delta\zeta_i)^2 = \frac{\delta r_i^2}{(n_i + 1)^2} + \frac{\zeta_i^2[\delta(n_i + 1)]^2 r_i^2}{(n_i + 1)^2} - \frac{2\zeta_i\delta n_i[\delta(n_i + 1)]}{(n_i + 1)^2}, \quad (7)$$

and

$$\frac{\zeta_i^2}{(\delta\zeta_i)^2} = \frac{r_i^2}{\delta r_i^2} + \frac{(n_i + 1)^2}{[\delta(n_i + 1)]^2} - \frac{r_i(n_i + 1)}{2\delta r_i[\delta(n_i + 1)]}. \quad (8)$$

By applying a similar approach for T_v we can obtain a similar expression for the s/n associated with estimating τ_i . Starting from Equation 5 we can show that,

$$\frac{\tau_i}{(\delta\tau_i)} = \left[\frac{q_i^2}{\delta q_i^2} + \frac{(t_i + 1)^2}{[\delta(t_i + 1)]^2} - \frac{q_i(t_i + 1)}{2\delta q_i[\delta(t_i + 1)]} \right]^{1/2}. \quad (9)$$

5 IMPLEMENTATION

5.1 Establishing s/n Thresholds

With our expressions for the s/n of ζ_i and τ_i we now explore the way in which the concept of an s/n threshold, beyond which the T_c and T_v statistics ‘flat-line’, may be integrated into our code for computing these statistics for a given survey. First we recall a fundamental property of both estimators, for a given m_* : by their construction, both T_c and T_v should have a Gaussian sampling distribution with mean zero and a variance equal to unity. We can therefore use the s/n expressions derived in the previous section to establish minimum s/n thresholds that will ensure the sampling distribution of both T_c and T_v is indeed Gaussian, with the correct mean and

variance, for each m_* – and particularly for fainter trial magnitudes closer to m_{lim}^f .

A procedure by which we can achieve this is illustrated in Figure 5 and the discussion which follows. In all three plots we present the following:

- Top panel: the T_c and T_v curves, shown as solid and dashed lines respectively, for a fixed, target value of δZ and δM respectively. We also indicate the imposed bright and faint apparent magnitude limits, m_{lim}^b and m_{lim}^f respectively.
- 2nd panel: here the achieved maximum (or peak) value of both δZ and δM , at each m_* , is shown in green, while the mean value of δZ and δM is shown in blue.
- 3rd panel: here we show, for each m_* , the resulting peak s/n indicated by the green curve, while the mean s/n is indicated by the blue curve. In this case the solid lines represent the s/n for ζ_i whilst the dashed lines are for τ_i .
- 4th panel: here we show a histogram of the apparent magnitude distribution for the survey under consideration.

Let us consider the SDSS-ET survey, shown in the left-hand plot of Figure 5. Here we have applied the usual JTH07 method with a target width of δZ and $\delta M = 0.015$. We use the phrase ‘target width’ as this version of the method seeks to maximise the sampling range of apparent magnitude within the JTH07 approach. Thus, we have allowed δZ and δM widths *smaller* than the targeted, fixed value to be included in the calculation. Therefore, it becomes clear that for the initial increments of m_* where the distance between m_* and m_{lim}^b is small (coupled with low numbers of galaxies), δZ and δM will not reach the target width.

In the 2nd panel of SDSS-ET, we show the resulting maximum value, as well as the mean value, of δZ and δM that was achieved for each m_* . In particular the mean value curves are clearly seen to fall below the target width, as described above, for the initial increments of m_* .

The choice of this width was made so that the T_c and

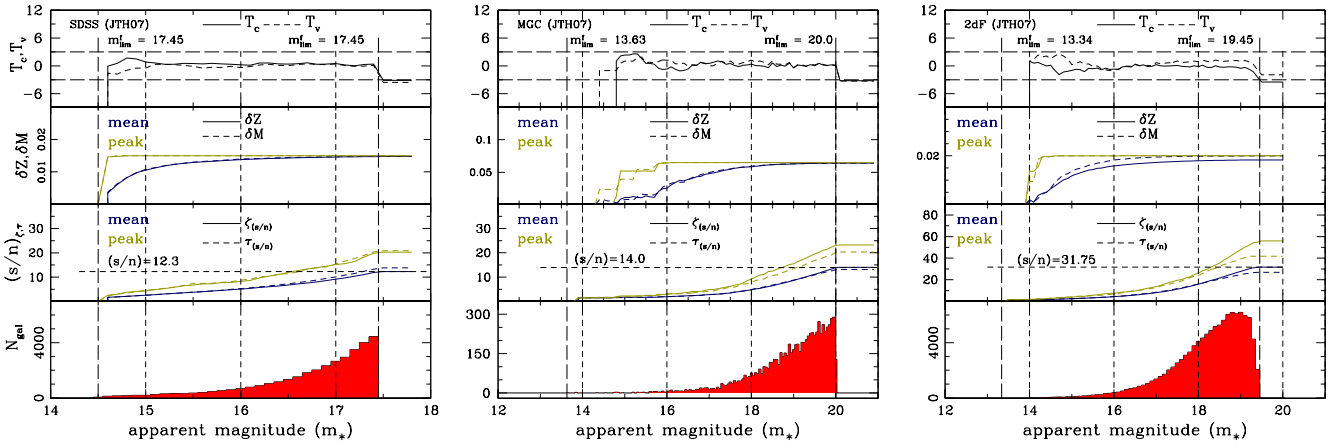


Figure 5. Plots demonstrating how we can establish a minimum signal-to-noise (s/n) threshold in SDSS-ET (left), MGC (middle) and 2dF (right) from the JTH07 approach. For each survey we show the imposed apparent magnitude limits, m_{lim}^b and m_{lim}^f , as vertical black dashed lines. In each case we choose a targeted, fixed value of δZ and δM (2nd panels) for T_c and T_v respectively (top panels) for which the resulting curves drop and ‘flat-line’ on or below the -3σ confidence limit of each test statistic. These widths correspond to δZ and $\delta M = 0.015$ (SDSS-ET), 0.065 (MGC) and 0.02 (2dF). As these are our target values, the 2nd panels for each survey show the resulting maximum (‘peak’ shown in green) value of δZ and δM that was achieved at each m_* , as well as the mean values (shown in blue). To maximise the sampling range of apparent magnitude with the JTH07 method, we have included where necessary δZ and δM widths smaller than the targeted, fixed value to be included in the calculation. Therefore, it becomes clear that for the initial increments of m_* where the distance between m_* and m_{lim}^b is small (coupled with low numbers of galaxies), δZ and δM will not reach the target width. The mean values of δZ and δM (shown in blue in the 2nd panels) for all surveys clearly illustrate this effect. In the case of MGC, in particular, we observe initial increments of the peak curves indicating that for no galaxy are we able to construct a separable region with δZ and $\delta M = 0.065$. By choosing the δZ and δM widths to drop to the -3σ limit beyond the m_{lim}^f , we have, in effect, established a minimum s/n threshold that should ensure our estimators are not subject to the effects of very sparse sampling at brighter trial apparent magnitudes. We find this choice corresponds to a $s/n \sim 12.3$ for the SDSS data, ~ 14.0 for MGC and ~ 31.75 for the 2dF.

T_v curves drop to on or below their -3σ confidence level at $m_{\text{lim}}^f = 17.45$. In the case of SDSS-ET this value of δZ and δM corresponds to a s/n level ~ 12.3 for both T_c and T_v . For this survey, therefore, one would need to maintain a minimum s/n threshold ~ 12.3 to ensure that the T_c and T_v statistics do not ‘flat-line’ due to very sparse sampling at magnitudes brighter than $m_{\text{lim}}^f = 17.45$. We will explore further the consequences of this in the following section.

In the remaining plots in Figure 5 we apply the corresponding procedure, with the same goal of ensuring that the ‘flat-lining’ behaviour occurs for sufficiently small values of the test statistics, to the MGC and 2dFGRS. For MGC in the middle plot, we require to set δZ and $\delta M = 0.065$ and find a mean $s/n \sim 14.0$ threshold. Finally, for 2dFGRS, we require to set δZ and $\delta M = 0.02$, which corresponds to a mean threshold of $s/n \sim 31.75$.

5.2 Imposing the s/n Thresholds

We can now use our pre-determined s/n levels, established in the previous section, and explore their impact on the T_c and T_v estimators. In Figure 6 we once again show the three surveys used for illustrative purposes in the same format as shown in Figure 5 and detailed in § 5.1.

Let us first consider the MGC data shown in the middle plot. As a simplistic approach to implementing an s/n threshold we have decided to keep the average s/n constant throughout the sampling procedure. This is achieved by keeping constant the number of galaxies counted in $S_1 \cup S_2$ (for ζ) and $S_3 \cup S_4$ (for τ). For MGC, to achieve the minimum s/n level of ~ 14.0 , already established from Figure 5, requires that the number of galaxies is equal to 150 in these combined regions. If we look at the 2nd panel for MGC we can observe the consequences for both δZ and δM as m_* increases towards the true magnitude limit, m_{lim}^f of the survey. Initially, we

see that δZ and δM are required to be rather large in size in order to achieve the minimum s/n level. This behaviour is expected and echoed by the histogram shown in the bottom panel of the plot. As the density of galaxies increases for fainter values of m_* we see a sharp decline in the required width of δZ and δM to achieve the same s/n . We also note that imposing a minimum s/n level restricts the magnitude sampling range within which T_c and T_v can reliably test completeness, particularly for brighter apparent magnitudes, and effectively introduces a value of m_* at which the test statistics ‘initialise’. In MGC, this initialisation occurs at an $m_* \sim 17.6$ mag.

If we now turn our attention to SDSS-ET on the left plot of Figure 6 we can see that the distribution of galaxies on the M - Z plane is such that we do not throw away much information on bright end of the apparent magnitude range. Both T_c and T_v initialise at around $m_* \sim 15.1$, after which we see a similar, steep drop in δZ and δM as was apparent with MGC. To achieve the minimum s/n level of ~ 12.3 the number of galaxies to be counted in separable regions is required to be 130 galaxies.

It is interesting to note that with the SDSS-ET survey, the T_c and T_v statistics initially fluctuate below -3σ between $15.1 \lesssim m_* \lesssim 15.5$. Similar behaviour is also observed with 2dF (see the right-hand plot). We recall that both SDSS-ET and 2dF surveys are well described by a bright and faint apparent magnitude limit, and as such are subject to natural restrictions of the maximum size of the ζ and τ sample regions that retain the separability assumptions of the estimators - see § 3 for further clarification of this point. In our implementation of an s/n threshold to our code, we have in this instance, allowed δZ and δM to grow in size beyond the limit imposed at the bright end. Therefore, until δZ and δM narrow to a width that defines the separable region within the survey limits, the estimators will indicate incompleteness. As we have already discussed, MGC can be well described by a faint limit only and is therefore not adversely affected by large values of δZ and δM .

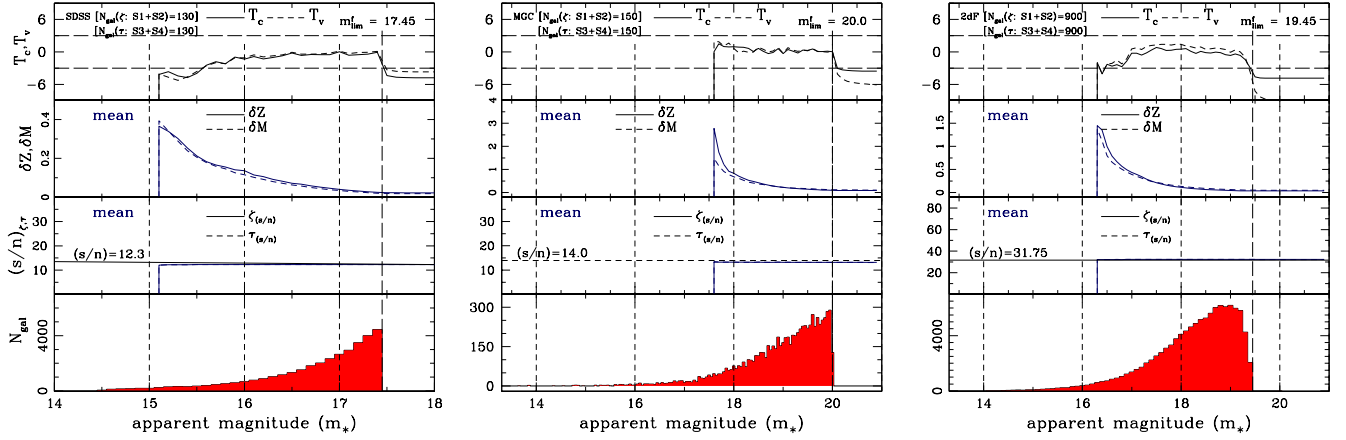


Figure 6. This figure shows the resulting T_c and T_v curves for all three surveys when we adopt a constant s/n level based on the approximate thresholds as established in Figure 5. The first distinguishing feature from that presented in Figure 5 is the deliberate omission of an imposed bright limit, m_{lim}^b . Secondly, in this simplified approach we can keep the average s/n constant by keeping constant the number of galaxies we count in the $S_1 \cup S_2$ (for ζ) and $S_3 \cup S_4$ for (τ) . By doing this, however, of course we sacrifice information at the *bright* end of the luminosity function, where the survey is too sparsely sampled to be included in the calculation. This effect is particularly obvious in the T_c and T_v results for MGC (middle plot) and 2dFGRS (right plot) and is mirrored in their respective histogram distributions. However, the advantage of this approach is that we now have *adaptive* δZ and δM widths for the respective estimators where, as we increase in m_* , the M - Z distribution is becoming more densely populated resulting in an almost asymptotic drop in the required widths to achieve the same s/n level. As already mentioned, we have omitted m_{lim}^b . This allows us to achieve the minimum s/n for a greater range in apparent magnitude, and therefore allows δZ and δM to grow as large as required. Such large values are only evident for initial values of m_* as show in the 2nd panels of each survey. For MGC (middle plot) this has no adverse effect on the respective T_c and T_v curves as MGC is equally well described by a single faint apparent magnitude limit, m_{lim}^f . However, we can see that for SDSS-ET and 2dF, the T_c and T_v statistics show initial fluctuations below -3σ which is to be expected since both surveys are defined with bright magnitude limits. As δZ and δM decreases in size with increased m_* , so the estimators become less sensitive to presence of m_{lim}^b .

Finally, with the 2dF survey on right-hand panel of Figure 6, we have set the number of galaxies to 900 which seems to satisfy our s/n criterion in our new scenario. As we have just discussed, there are slight fluctuations below -3σ at bright values of m_* , i.e. for $m_* \sim 16.4$ mag. These correspond to the adoption of large widths for δZ and δM . It should be noted that even with the *minimum* s/n value, one can anticipate the true faint limit, m_{lim}^f , of 2dF being identified as brighter than the published limit of $m_{\text{lim}}^f = 19.45$ if one were to move to higher s/n levels.

6 DISCUSSION AND CONCLUDING REMARKS

In this article we have introduced a method which attempts to optimize the completeness estimators, suitable for application to double-truncated galaxy survey data, as previously developed by Johnston *et al.* (2007). Our new approach resembles an “adaptive smoothing” procedure which seeks to maintain a constant level of ‘information’ – as characterised by the signal-to-noise ratio computed for our test statistics – allocated to each galaxy in the survey. In applying this methodology to three well understood and characterized surveys, we have demonstrated the importance of properly accounting for the impact of sparse sampling in each galaxy survey. Furthermore, our results indicate that – without adopting such a procedure – the testing of magnitude completeness may be compromised, and spurious values for the ‘true’ apparent magnitude limit(s) may be inferred. Thus, sparse sampling effects may impact adversely on previous applications of product-limit estimators which have been carried out in the literature to doubly-truncated data sets e.g. Efron and Petrosian (1999).

The current article is the first of a two-part story. In the current paper we have set out to optimise our completeness estimators

by imposing a lower limit on the number of galaxies contained in (and hence a lower limit on the width of) the rectangular regions we identify in the M - Z distribution of our data. This lower limit ensures that the Gaussian sampling distribution, with mean zero and variance unity, of our T_c and T_v statistics is preserved over the range of m_* where the optimization is possible. In an upcoming publication (Johnston *et al.* 2010, in preparation) we will consider in more detail the practical implementation of these optimised estimators – and in particular how we may use them to assign error bars to T_c and T_v , and hence to compute confidence limits for the faint apparent magnitude limit, m_{lim}^f , properly accounting for the correlations in T_c and T_v between neighbouring values of the trial magnitude limit m_* .

ACKNOWLEDGEMENTS

RJ would like to thank David Valls-Gabaud for his insightful comments and also the funding bodies EPSRC (UK) and the National Research Foundation (South Africa).

The Millennium Galaxy Catalogue consists of imaging data from the Isaac Newton Telescope and spectroscopic data from the Anglo Australian Telescope, the ANU 2.3m, the ESO New Technology Telescope, the Telescopio Nazionale Galileo and the Gemini North Telescope. The survey has been supported through grants from the Particle Physics and Astronomy Research Council (UK) and the Australian Research Council (AUS). The data and data products are publicly available from <http://www.eso.org/~jliske/mgc/> or on request from J. Liske or S.P. Driver.

The SDSS-ET data-set was kindly provided by Mariangela Bernardi and Ravi Sheth. Funding for the creation and distribu-

tion of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is <http://www.sdss.org>. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are the University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, the Johns Hopkins University, Los Alamos National Laboratory, the Max Planck Institute for Astronomy (MPIA), the Max Planck Institute for Astrophysics (MPA), New Mexico State University, the University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

- Bernardi, M. *et al.* : 2003, *A J* **125**, 1817
 Colless, M., *et al.* : 2001, *MNRAS* **328**, 1039
 Cross, N. J. G., Driver, S. P., Liske, J., Lemon, D. J., Peacock, J. A., Cole, S., Norberg, P., and Sutherland, W. J.: 2004, *MNRAS* **349**, 576
 Drinkwater, M. J. *et al.* : 2010, *MNRAS* **401**, 1429
 Efron, B. and Petrosian, V.: 1992, *Astrophys. J.* **399**, 345
 Efron, B. and Petrosian, V.: 1999, *J. Amer. Statist. Assoc.* **94**(447), 824
 Hubble, E. P.: 1926, *Astrophys. J.* **64**, 321
 Hudson, M. J. and Lynden-Bell, D.: 1991, *MNRAS* **252**, 219
 Johnston, R., Teodoro, L., and Hendry, M.: 2007, *MNRAS* **376**, 1757
 Lilly, S. J., *et al.* : 2009, *ApJ Suppl.* **184**, 218
 Liske, J., Lemon, D. J., Driver, S. P., Cross, N. J. G., and Couch, W. J.: 2003, *MNRAS* **344**, 307
 Lynden-Bell, D.: 1971, *MNRAS* **155**, 95
 Rauzy, S.: 2001, *MNRAS* **324**, 51
 Schmidt, M.: 1968, *Astrophys. J.* **151**, 393
 Zucca, E., Bardelli, S., Bolzonella, M., Zamorani, G., Ilbert, O., Pozzetti, L., Mignoli, M., Kovač, K., Lilly, S., Tresse, L., Tasca, L., Cassata, P., Halliday, C., Vergani, D., Caputi, K., Carollo, C. M., Contini, T., Kneib, J., Le Fèvre, O., Mainieri, V., Renzini, A., Scodreggio, M., Bongiorno, A., Coppa, G., Cucciati, O., de La Torre, S., de Ravel, L., Franzetti, P., Garilli, B., Iovino, A., Kampczyk, P., Knobel, C., Lamareille, F., Le Borgne, J., Le Brun, V., Maier, C., Pellò, R., Peng, Y., Perez-Montero, E., Ricciardelli, E., Silverman, J. D., Tanaka, M., Abbas, U., Bottini, D., Cappi, A., Cimatti, A., Guzzo, L., Koekemoer, A. M., Leauthaud, A., Maccagni, D., Marinoni, C., McCracken, H. J., Memeo, P., Meneux, B., Moresco, M., Oesch, P., Porciani, C., Scaramella, R., Arnouts, S., Aussel, H., Capak, P., Kartaltepe, J., Salvato, M., Sanders, D., Scoville, N., Taniguchi, Y., and Thompson, D.: 2009, *A&A* **508**, 1217